DOCUMENT RESUME

ED 331 879 TM 016 428

AUTHOR Hambleton, Ronald K.; And Others
TITLE Influence of Item Parameter Errors in Test Development.
SPONS AGENCY Graduate Management Admission Council, Princeton, NJ.
PUB DATE Aug 90
NOTE 19p.; Paper presented at the Annual Meeting of the American Psychological Association (98th, Boston, MA, August 10-14, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Error of Measurement; *Estimation (Mathematics); Item Response Theory; *Sampling; *Simulation; *Test Construction; Testing Problems; Test Items
IDENTIFIERS *Information Function (Tests); *Item Parameters

ABSTRACT

Item response theory (IRT) model parameter estimates have considerable merit and open up new directions for test development, but misleading results are often obtained because of errors in the item parameter estimates. The problem of the effects of item parameter estimation errors on the test development process is discussed, and the seriousness of the problem is demonstrated with simulated data sets. Solutions are offered for this problem in test development practice, which arises because item information functions are determined by item parameter values that in turn contain error. When the best items are selected on the basis of their statistical characteristics, there is a tendency to capitalize on chance due to errors in the item parameter estimates; among the generally promising test items, items with parameter estimates that are the most overestimated are also the most likely to be selected. As a result, the test falls short of the test desired or expected. Simulation studies using a hypothetical pool of 150 test items with sample sizes of 1,000 and 400 confirmed that tests do not perform as well as expected when items are selected to match a target test information function and standard errors are correspondingly underestimated. The following suggestions for eliminating this problem are presented: (1) use large samples in item calibration to gain precision in item parameter estimates; (2) revise the extreme item parameter estimates by subtracting one or two standard errors from their values; and (3) exceed the desired target information by 20 to 30%. Two tables and six graphs supplement the discussion. (SLD)

Influence of Item Parameter Errors in Test Development

Ronald K. Hambleton, Russell W. Jones
University of Massachusetts at Amherst

and

H. Jane Rogers
Columbia Teachers College

Abstract

Item response models are finding increasing use in achievement and aptitude test development. This particular item response theory (IRT) application involves the selection of test items via a consideration of their item information functions. But a problem arises because item information functions are determined by item parameter values which in turn contain error. When the "best" items are selected on the basis of their statistical characteristics, there is a tendency to "capitalize on chance" due to errors in the item parameter estimates: among the generally promising test items, items with parameter estimates that are the most overestimated are also the most likely to be selected. As a result, the test falls short of the test that was desired or expected. The purposes of this paper are (1) to highlight this problem, which to date has received almost no attention in the IRT literature, (2) to demonstrate the seriousness of the problem with several simulated datasets, and (3) to offer a conservative solution for addressing the problem in IRT-based test development.

Influence of Item Parameter Errors in Test Development[1,2,3]

Ronald K. Hambleton, Russell W. Jones
University of Massachusetts at Amherst

and

H. Jane Rogers
Columbia Teachers College

Over the last 20 years, many test developers have begun to use item response theory (IRT) models and methods rather than classical measurement models in their test development and related technical work (Hambleton, 1989; Hambleton & Swaminathan, 1985; Lord, 1980). Item response theory, particularly as reflected in the one-, two-, and three-parameter logistic models for dichotomously scored items, is receiving increasing attention from test developers in test design and test item selection, in addressing item bias, in computer-administered testing, and in the equating and reporting of test scores. Nearly all major test publishers, state departments of education, and large school districts currently use IRT models in some capacity in their testing work.

One problem that arises when applying IRT models in test development involves "capitalizing on chance" due to positive errors in some item parameter estimates. The problem arises because test developers, not surprisingly, prefer to select test items with the highest discrimination indices. But high discrimination indices, on the average, are spuriously

---

[1] The research described in the paper was funded by the Graduate Management Admission Council (the Council). The Council encourages researchers to formulate and freely express their own opinions. The opinions here are not necessarily those of the Council.

[2] Laboratory of Psychometric and Evaluative Research Report No. 208. Amherst, MA: University of Massachusetts, School of Education.

[3] Paper presented at the annual meeting of APA, Boston, 1990.

high because of positive errors in item parameter estimation. As a result, tests often fall short, statistically, of what is expected, and standard errors associated with ability estimates are underestimated (if the inflated item parameter estimates are used) which leads to overconfidence in the ability estimates when the bands are set. Whether or not additional problems are created, such as bias in ability estimation, is not known. It is worth noting that an analogous problem arises in item selection in computerized adaptive testing.

The purposes of this paper are (1) to highlight the problem of item parameter estimation errors on the test development process, which to date has received almost no attention in the IRT literature, (2) to demonstrate the seriousness of the problem with several simulated datasets, and (3) to offer a conservative solution for addressing the problem in test development practice.

Prior to addressing the three purposes of the paper, a brief introduction will be offered to item and test information functions and the ways in which these functions are used in IRT test development.

Item and Test Information Functions

Item response models provide a powerful method of describing items and tests and selecting test items. The method involves the use of item information functions. Item information functions play an important role in test development in that they display the contribution items make to ability estimation at points along the ability continuum. This contribution depends to a great extent on an item's discriminating power (the higher it is, the steeper the slope of the item characteristic curve and the more information the item provides). The location on the ability

continuum at which this contribution is realized is dependent on the item's difficulty (Hambleton, 1989).

The test information function for a test, reported at each ability level, is simply the sum of the item information functions. Thus, the contribution of individual test items can be determined without knowledge of other items in the test. The amount of information that a particular test provides at an ability level influences the precision with which ability is measured -- the more information the more accurate the ability estimates.

Lord (1980) outlined a procedure for the use of item information functions to build tests to meet any desired set of statistical specifications. The procedure employs an item bank with item statistics available for the IRT model of choice, with accompanying item information functions. The procedure consists of the following steps:

1. Decide on the shape of the desired test information function (called the target information function).
2. Select items from the item bank with item information functions that fill up the hard-to-fill areas under the target information function.
3. After each item is added to the test, calculate the test information function for the selected test items.
4. Continue selecting items until the test information function approximates the target information function to a satisfactory degree.

The four steps above were used in the research described below.

Item Parameter Estimation Errors

In building tests to fit target test information functions, test developers will tend to choose the most discriminating items that also satisfy the necessary content specifications. For a given test length, such a procedure leads to a test with maximum information. The problem is that high item discrimination indices, on the average, tend to be over-estimated. (In a similar fashion, low item discrimination indices, on the average, tend to be under-estimated, but this is not a problem because test developers rarely have interest in selecting items with low discrimination indices.) This is the well-known problem of "regression effects due to errors of measurement" which is usually discussed and considered in the context of test score interpretations. But, it occurs with item parameter estimates too -- high item parameter estimates tend to be over-estimated, and low item parameter estimates tend to be under-estimated. The amount of error will depend on sample size: large samples lead to small errors; smaller samples lead to large errors due to the regression effect.

The consequence of errors in the item parameter estimates is that often tests do not function as well in practice as test developers expect based upon a consideration of the item parameter estimates for items in their bank. Test developers tend to choose the most discriminating items available to them but the true item discrimination values for many of these items are somewhat lower than their estimated values. Hence, the test does not function as well as might be expected. To demonstrate this point, under several conditions, a simulation study was carried out.

A hypothetical pool of 150 test items to fit the two-parameter model was produced using a computer program prepared by Hambleton and Rovinelli (1973). To facilitate the interpretation of results, all items were taken

to have a true discrimination parameter equal to one. Then, ability scores (normally distributed, 0,1) for 1000 examinees along with simulated examinee item responses were generated and item parameter estimates (difficulty and discrimination) were obtained. A second sample of examinees was drawn from the same examinee population and the item parameter estimates were obtained again. Finally, both analyses were repeated using examinee sample sizes of 400.

A target information function was then specified, and the "best" 25 items to provide the desired test were selected using the item parameter estimates obtained from the first samples drawn (N=1000, N=400). Tests constructed in this way were capitalizing on the positive errors in some of the item parameter estimates. Tables 1 and 2 present the parameter

---

Insert Tables 1 and 2 about here

---

estimates for items in each test (which were constructed using the first sample estimates) and the parameter estimates for the same 25 items obtained in the second samples. Unlike the first sample estimates, the second sample estimates would not be biased in the sense of being over-estimated. What is very evident from both tables is that selected test items had item discrimination estimates which were too high (in relation to their true values), whereas, in the second samples, a more random pattern (15 of 25 in each sample [N=1000; N=400] were too high, and 10 of 25 were too low in relation to their true values) was observed. The results clearly show the effects of "capitalizing on chance." A comparison of the a-parameter estimates in Tables 1 and 2 also shows the role of sample size on the extent of over-estimation. The problem is worse with the smaller sample sizes ($\bar{a}$ = 1.18 in the large sample; $\bar{a}$ = 1.26 in the small sample).

Graphical displays of the test information functions from the estimated and true parameter values with samples one and two and for large (N = 1000) and small (N = 400) sample sizes are shown in Figures 1 to 4. Figures 1 and 3 highlight the size of the bias that might be expected in the test information function when the most discriminating items in the bank are selected. The bias is substantial, and, of course, largest when N = 400. The actual 25-item test would need to be lengthened by 30 to 50% to produce a test to match the desired test over the main portion [-2, +2] of the ability scale (see Figure 6). With the larger sample, N = 1000, the actual test would need to be lengthened by 20% to 30% to produce a test to match the desired test over the same interval (see Figure 5). Figures 2 and 4 highlight the comparison of true and estimated test information functions when unbiased item parameter estimates are used in the calculations. The smaller difference between the curves in Figure 2 than Figure 4 is due to the use of a larger sample size in item parameter estimation, and consequently more accurately estimated item parameters.

Insert Figures 1 to 6 about here

There are several implications for practice: (1) tests do <u>not</u> perform as well as expected when the "best" items are selected to match a target test information function, and (2) standard errors are, correspondingly, under-estimated (assuming that the first set of values is taken as "true values") and so over-confidence in ability scores will result. These results provide rather dramatic evidence of the influence of selecting the "best" items from an item bank to make up a test. In future simulations, we will also investigate relationships among item bank size, test length, and sample sizes used in calibrating the test items. In general, we expect

to find that the larger the bank, the shorter the test, and the smaller the sample size, the more serious the regression problem will be.

At least three steps can be taken to reduce the problem:

1. Use large samples in item calibration to gain precision in item parameter estimates. An increase in the precision of item parameter estimates will reduce the significance of the regression effect due to errors of measurement.
2. Revise the extreme item parameter estimates by subtracting one or two standard errors from their values.
3. Depending on the sample size used in item parameter estimation (and assuming suggestion 2 has not been implemented), exceed the desired target information function by 20% to 30%.

If one or more of the above suggestions are implemented, the problem associated with using over-estimated item parameters in ability and standard error estimation can be minimized.

Conclusion

There is ample evidence in the psychometric literature to support the expanded use of IRT models in test development and analyses (see, for example, Green, Yen, & Burket, 1989). The main point of this paper is that IRT model parameter estimates have considerable merit and open up new directions for test development, but misleading results will often be obtained because of errors in the item parameter estimates. Test developers must strive for large-sized samples in estimating item parameters and exceed the desired target information curve by (perhaps) 20% to 30% to correct for capitalizing on chance in item selection. Alternately, item parameter estimates which are substantially above the mean parameter values could be reduced by one or two standard errors.

There are many IRT methodological issues that must be attended to by test developers, but the one described in this paper seems particularly relevant for improving IRT test development practice.

References

Green, D. R., Yen, W. M., & Burket, G. R. (1989). Experiences in the application of item response theory in test construction. Applied Measurement in Education, 2(4), 297-312.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. Linn (Ed.), Educational measurement (3rd edition, pp. 147-200). New York: Macmillan.

Hambleton, R. K., & Rovinelli, R. J. (1973). A Fortran IV program for generating examinee response data from logistic test models. Behavioral Science, 17, 73-74.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer Academic Publishers.

Lord, F. M. (1980). Application of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Table 1

Summary of Item Statistics
(N=1000)

Test Item	First Estimates b	First Estimates a	Second Estimates b	Second Estimates a	True Values b	True Values a
1	0.74	1.17	0.74	1.09	0.83	1.00
2	-0.36	1.16	-0.39	1.04	-0.35	1.00
3	-1.85	1.14	-1.97	0.89	-1.85	1.00
4	-0.70	1.17	-0.68	0.99	-0.66	1.00
5	1.31	1.15	1.32	1.12	1.35	1.00
6	1.63	1.17	1.76	1.05	1.77	1.00
7	-1.83	1.24	-1.86	1.08	-1.92	1.00
8	1.24	1.17	1.40	0.93	1.42	1.00
9	-0.21	1.14	-0.24	1.04	-0.25	1.00
10	-1.84	1.14	-1.92	0.96	-1.79	1.00
11	1.12	1.20	1.38	0.87	1.23	1.00
12	-0.58	1.18	-0.63	1.08	-0.61	1.00
13	1.60	1.30	1.90	1.04	1.94	1.00
14	0.53	1.23	0.47	0.98	0.60	1.00
15	-1.40	1.17	-1.67	0.85	-1.44	1.00
16	1.41	1.15	1.50	1.05	1.58	1.00
17	-1.08	1.20	-1.30	0.92	-1.17	1.00
18	0.75	1.22	1.03	0.81	0.88	1.00
19	1.15	1.18	1.25	1.02	1.30	1.00
20	-1.73	1.24	-1.81	1.06	-1.90	1.00
21	-0.17	1.14	-0.19	1.00	-0.20	1.00
22	1.32	1.14	1.36	0.93	1.38	1.00
23	0.29	1.19	0.39	1.12	0.41	1.00
24	-1.88	1.14	-1.90	1.03	-1.87	1.00
25	-0.09	1.23	-0.05	1.07	-0.03	1.00

Table 2

Summary of Item Statistics
(N=400)

Test Item	First Estimates b	First Estimates a	Second Estimates b	Second Estimates a	True Values b	True Values a
1	-1.52	1.21	-1.62	1.13	-1.54	1.00
2	0.71	1.45	0.59	1.08	0.83	1.00
3	1.16	1.37	1.60	0.87	1.50	1.00
4	-0.44	1.18	-0.64	0.88	-0.35	1.00
5	0.99	1.29	1.20	0.93	1.21	1.00
6	-0.78	1.29	-0.80	1.06	-0.75	1.00
7	-0.39	1.21	-0.25	1.18	-0.32	1.00
8	0.46	1.26	0.41	1.13	0.54	1.00
9	0.56	1.52	0.79	0.91	0.71	1.00
10	0.70	1.18	0.58	1.24	0.82	1.00
11	0.67	1.19	1.06	0.90	0.71	1.00
12	1.75	1.17	-1.27	1.26	1.94	1.00
13	0.08	1.48	0.72	1.01	0.25	1.00
14	0.01	1.18	2.08	0.68	0.15	1.00
15	0.55	1.25	1.02	1.04	0.60	1.00
16	-0.36	1.19	1.26	1.20	-0.38	1.00
17	0.52	1.19	1.60	0.87	0.64	1.00
18	0.01	1.20	1.20	0.93	0.07	1.00
19	1.58	1.23	-1.58	0.97	1.78	1.00
20	-0.84	1.21	1.10	1.16	-0.70	1.00
21	0.92	1.19	-1.03	1.02	1.01	1.00
22	-1.46	1.23	-1.68	0.99	-1.53	1.00
23	-0.11	1.20	-0.42	1.05	-0.20	1.00
24	1.01	1.18	-0.24	1.13	1.23	1.00
25	1.24	1.29	0.59	1.11	1.38	1.00

Figure 1. Test Information from Estimated and True Item Parameters

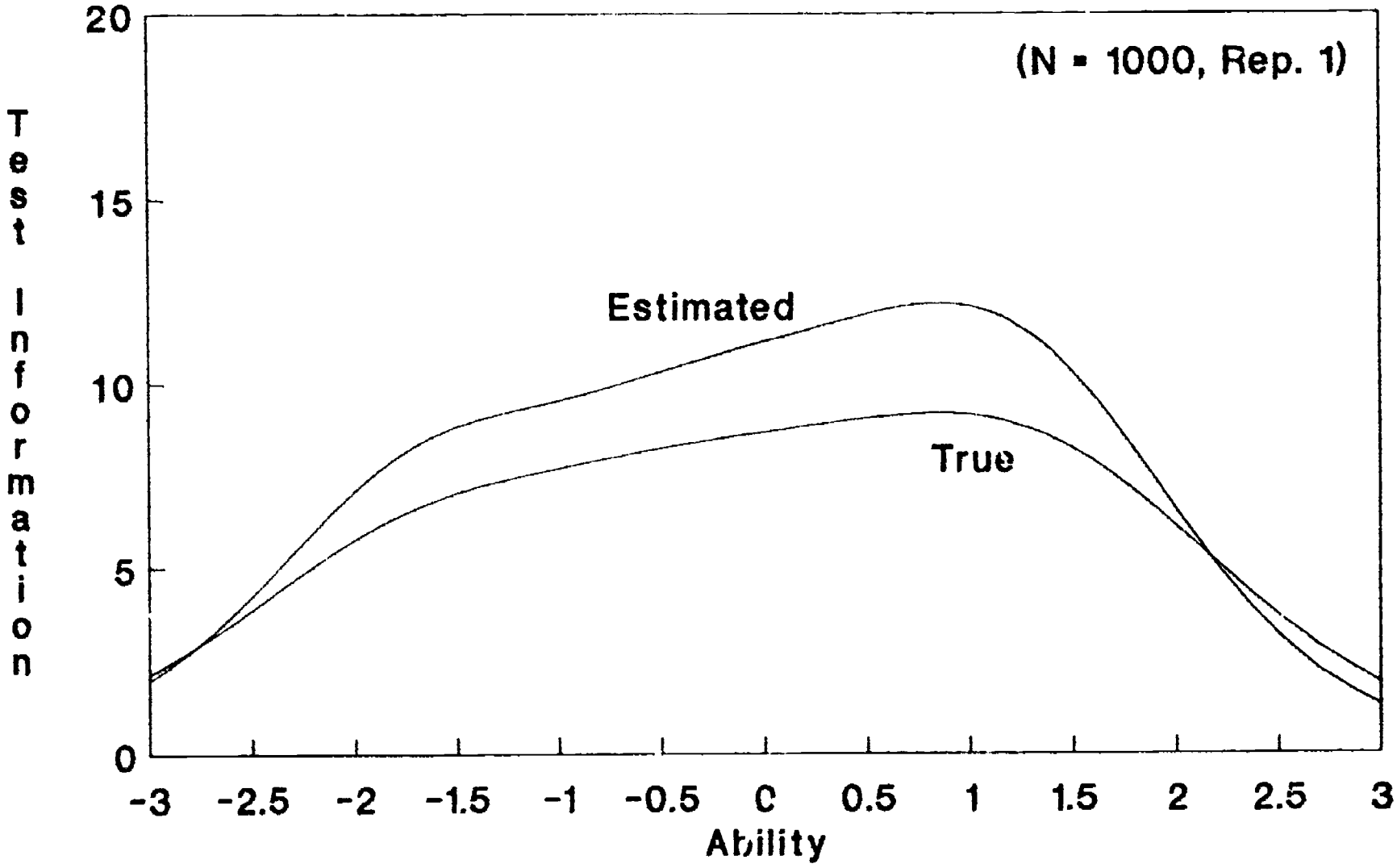

Figure 2. Test Information from Estimated and True Item Parameters

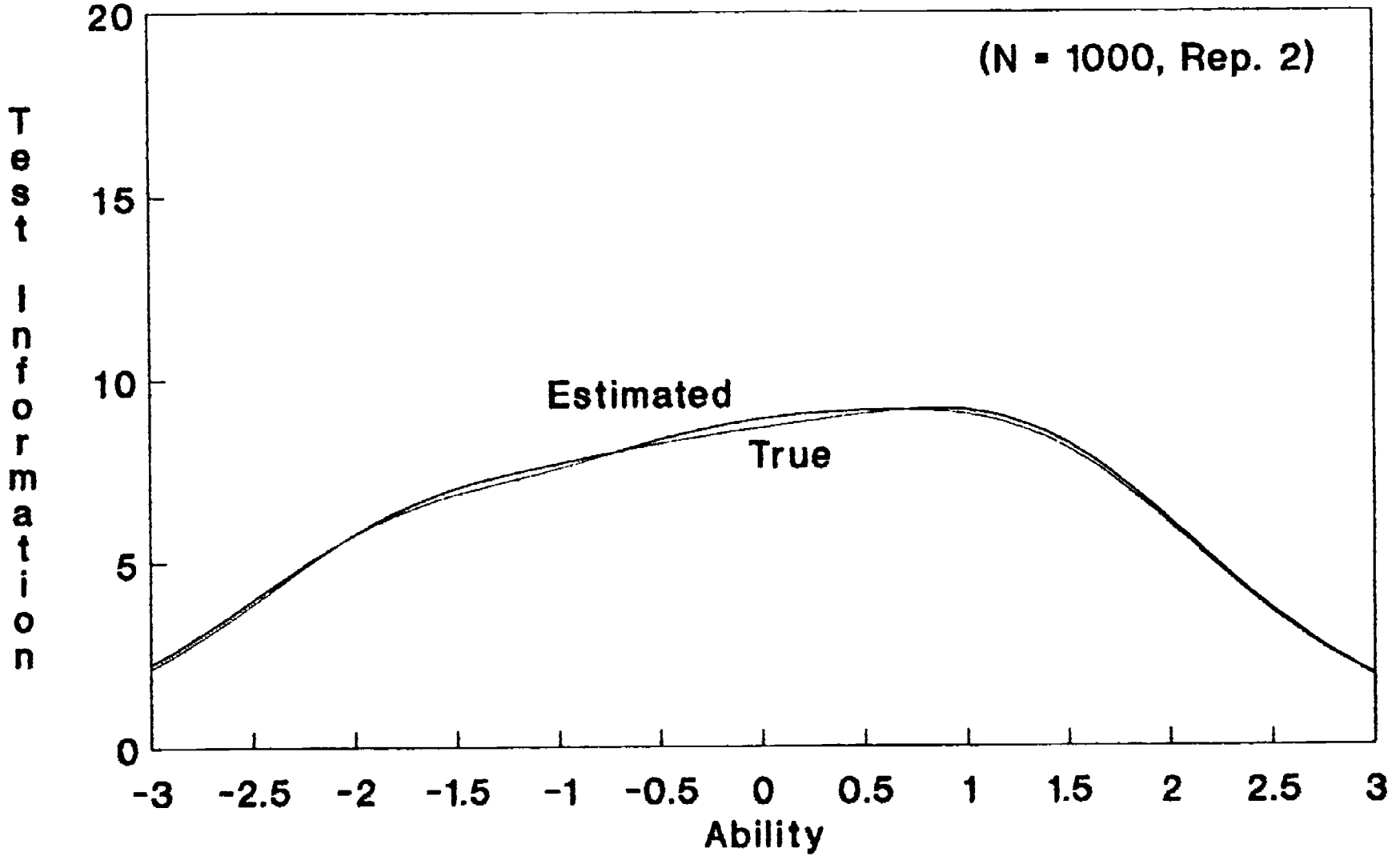

Figure 3. Test Information from Estimated and True Item Parameters

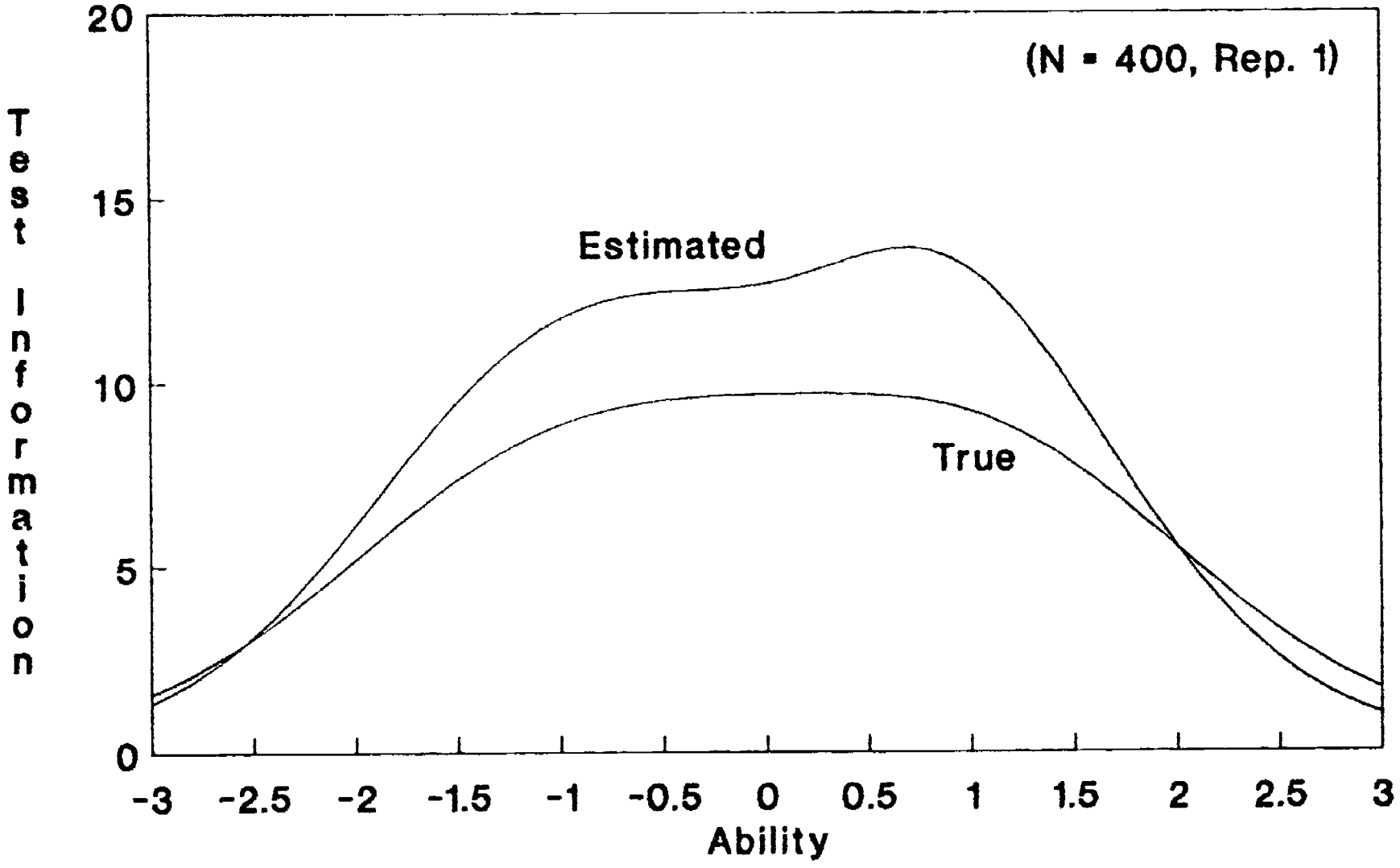

Figure 4. Test Information from Estimated and True Item Parameters

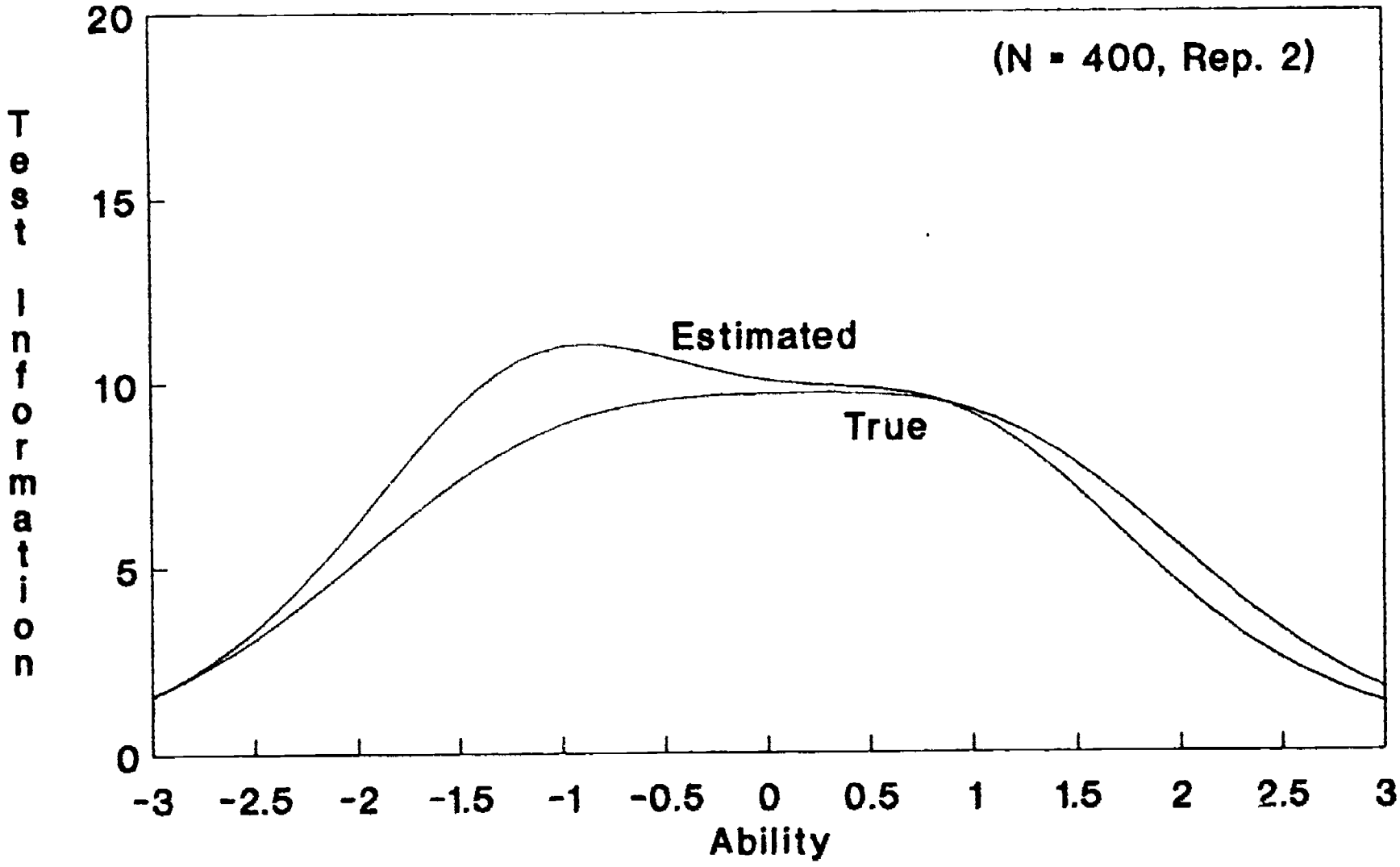

Figure 5. Efficiency Function for Estimated vs. True Item Parameters

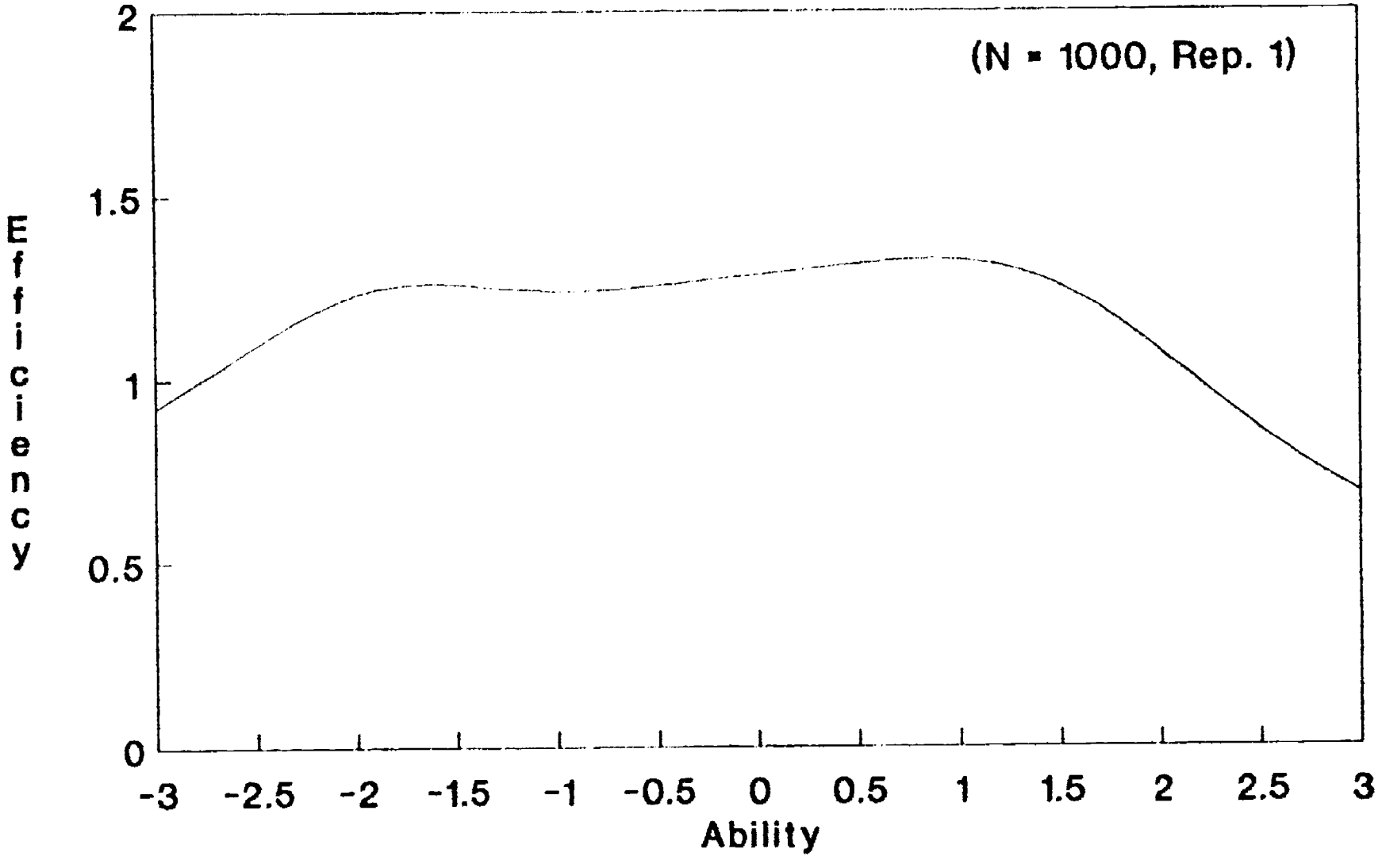

Figure 6. Efficiency Function for Estimated vs. True Item Parameters